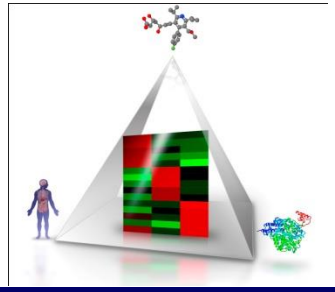
Artificial Intelligence in Drug Design – What is Realistic, What are Illusions?

Andreas Bender, PhD
Natural Philosopher for Molecular Informatics
Centre for Molecular Science Informatics
University of Cambridge

Associate Director CPSS DSAI, AstraZeneca
Fellow of King's College, Cambridge
Co-Founder, Healx Ltd.
Co-Founder, PharmEnable Ltd.

UNIVERSITY OF CAMBRIDGE



This is an academic talk

I am today speaking purely in my capacity as an academic

Everything presented is my own personal opinion and not that of any employer, funder or anyone else

Lots of things happening – time for critical evaluation of where we are

SPOTLIGHT · 30 MAY 2018

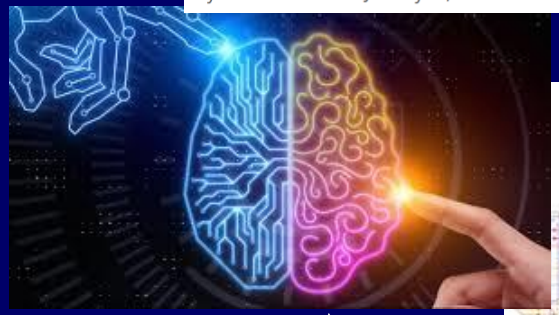
How artificial intelligence is changing drug discovery

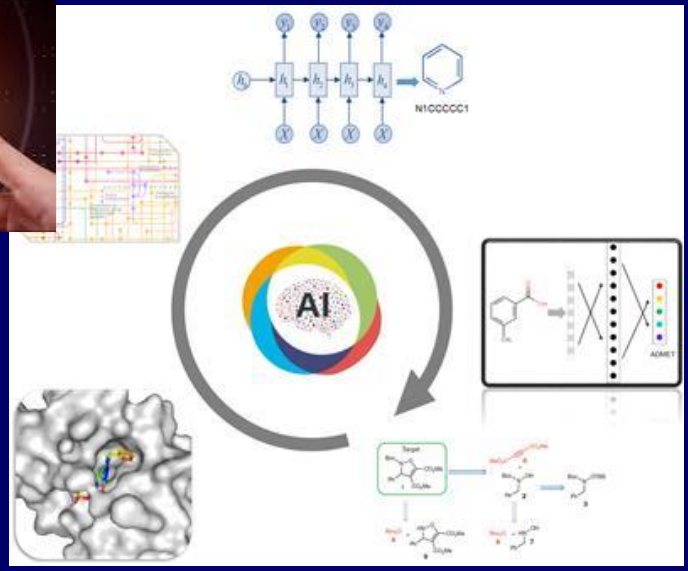
World first breakthrough in AI drug discovery

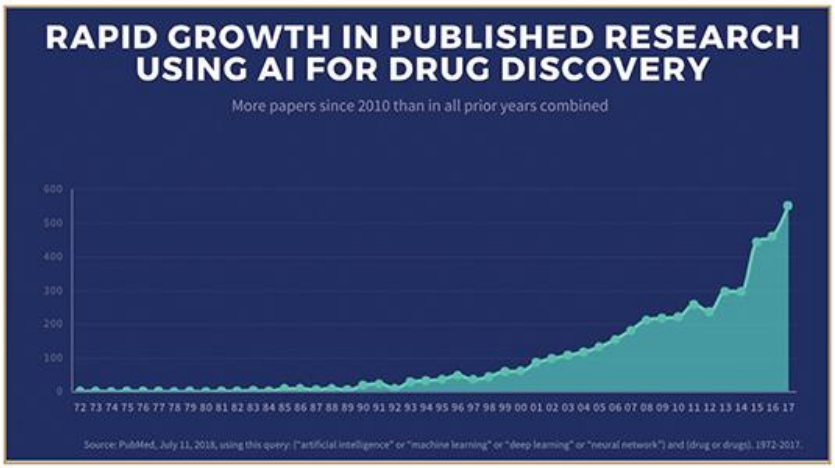
By **Emma Morriss** - January 30, 2020

AI 2020:

THE FUTURE OF DRUG DISCOVERY







Bzzzz…

Old enough to remember 2000 biotech bubble, Human Genome Project, etc.

T. Reiss, Trends in Biotechnology, 2001:

“The number of drug targets will increase by at least one order of magnitude and target validation will become a high-throughput process.”
“More drug targets… 3,000–10,000 targets compared with 483”

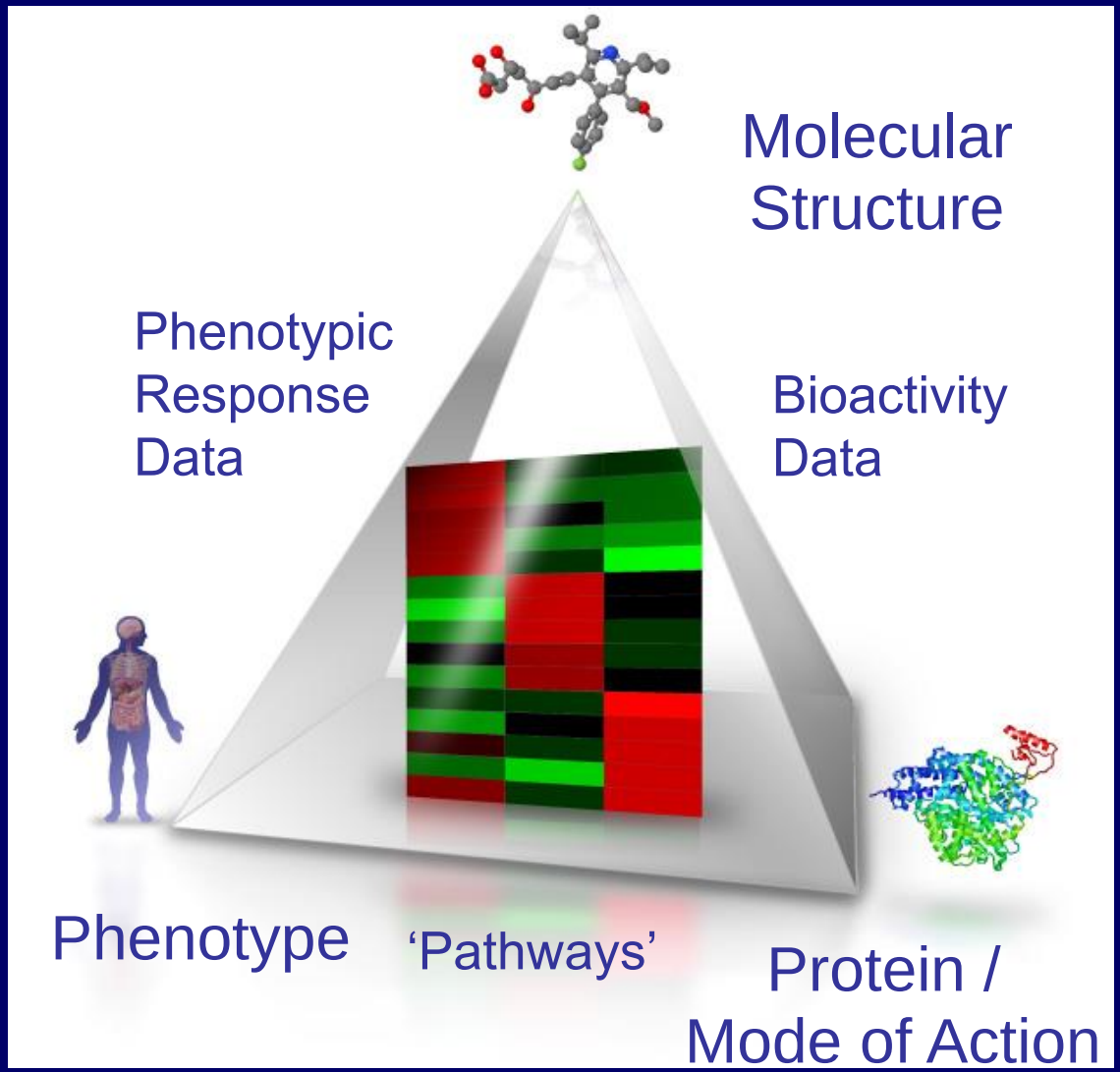
Recent (2017) estimates of drug targets put the number currently at around 667

http://www.drugdiscovery.net/2020/01/21/omics-data-so-where-is-the-signal-please/

The data landscape, deep learning, biology… and humans

- Chemical and biological data – in early discovery, vs later stages
- Easy labels, easy models
- Deep learning?
- Biology, biology, biology…
- The interface of AI, human psychology, and society

A simple view on the world: Linking Chemistry, Phenotype, Targets / Mode of Action



a.k.a. "The world is flat"

So what's the point of it all?
We would like to answer questions!

- "What is the reason upon treatment with A for phenotypic effect B?"
  -> *Mode of Action*
- "Which compound should I make to achieve effect C in a biological system?"
  -> *Chemistry*
- "Does patient D or patient E respond better to drug F?"
  -> *Phenotype / Phenotype Change*

BUT…The world is not flat. What now?

- Links between drugs/targets/diseases are quantitative (and incompletely characterized)

- Subtle differences in eg compound effects (partial vs full agonists, off-targets, residence times, etc.)

- Effects are state-dependent (variation between individuals, … even by what you have eaten in the morning…), often not captured in the data we have

- Phenotyping in particular is sparse, subjective (deep phenotyping as the answer?)

- We don’t properly understand biology (‘the system’), so we don’t know what to measure, label

Data depends on context: eg early discovery vs safety

Early discovery

- Often ‘simple’ readouts (eg activity on protein), hence…
- **Large number of data points** for training models

- ***Models have clear labels*** (within limits of model system; eg ‘ligand is active against protein at IC50<10uM’, logD, …)
- Good for model generation: *Many*, *clearly categorized* data points
- Less good for *in vivo* relevance

Later stage/safety

- Quantitative data (dose, exposure, …)
- More complex models (to generate data), and ***fuzzy labels*** (classes ‘depend’, on exposure, multiple eg histopathological endpoints) – hence…
- ***Less, and less clearly labelled data***: Difficult from machine learning angle

- Data: *Recording complex data in format suitable for mining* – eg animal data tricky, even within single company

Problem setting in early discovery vs safety

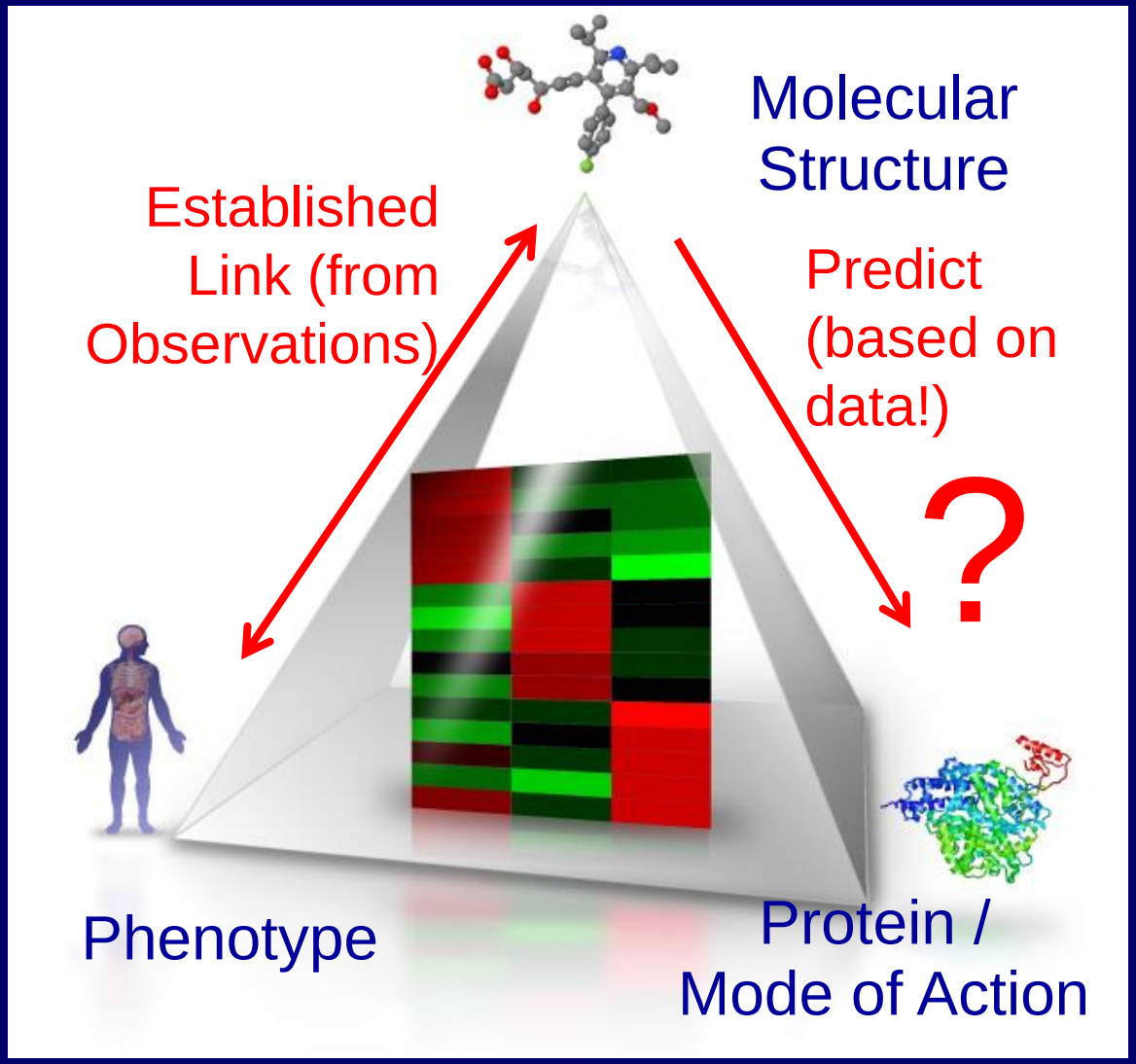
Early discovery

- Discovery setting – ‘find me suitable 100s or 1000s out of a million’ (eg screening)
- Anything fulfilling (limited) set of criteria will do ‘for now’, predicting *presence of something*
- Computationally *generative* models often useful

Late stage/safety

- Need to predict for *this particular data point*
- *Large number of criteria to rule out, based on limited data*… predicting *absence of ‘many things’* (eg different modes of toxicity)
- *Predictive* models (more tricky than generative; eg data coverage limiting)

Starting from *in vivo* efficacy we can hypothesize the MoA, based on ligand chemistry



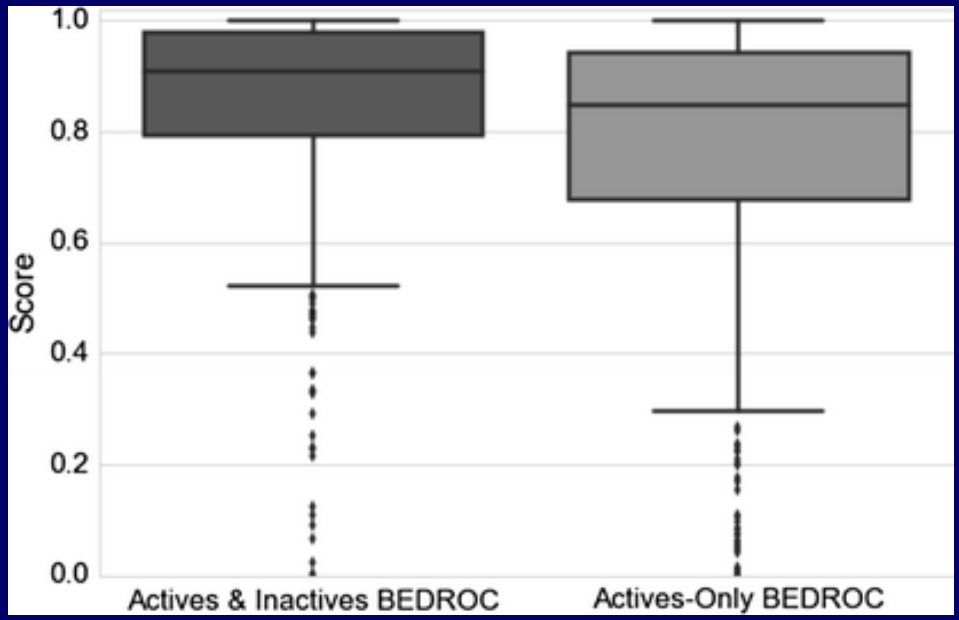
A. Koutsoukas *et* al., J Proteomics 2011 (74) 2554 – 2574.

Public target prediction model, based on ~200 mio data points

- Work of Lewis Mervin, with AstraZeneca
- 2015, *J. Cheminformatics* (7) 51
- ChEMBL actives (~300k), PubChem inactives (~200m)
- Can be retrained on in-house data
- 1,080 targets

- https://github.com/lhm30/PIDGIN

Also data publicly available



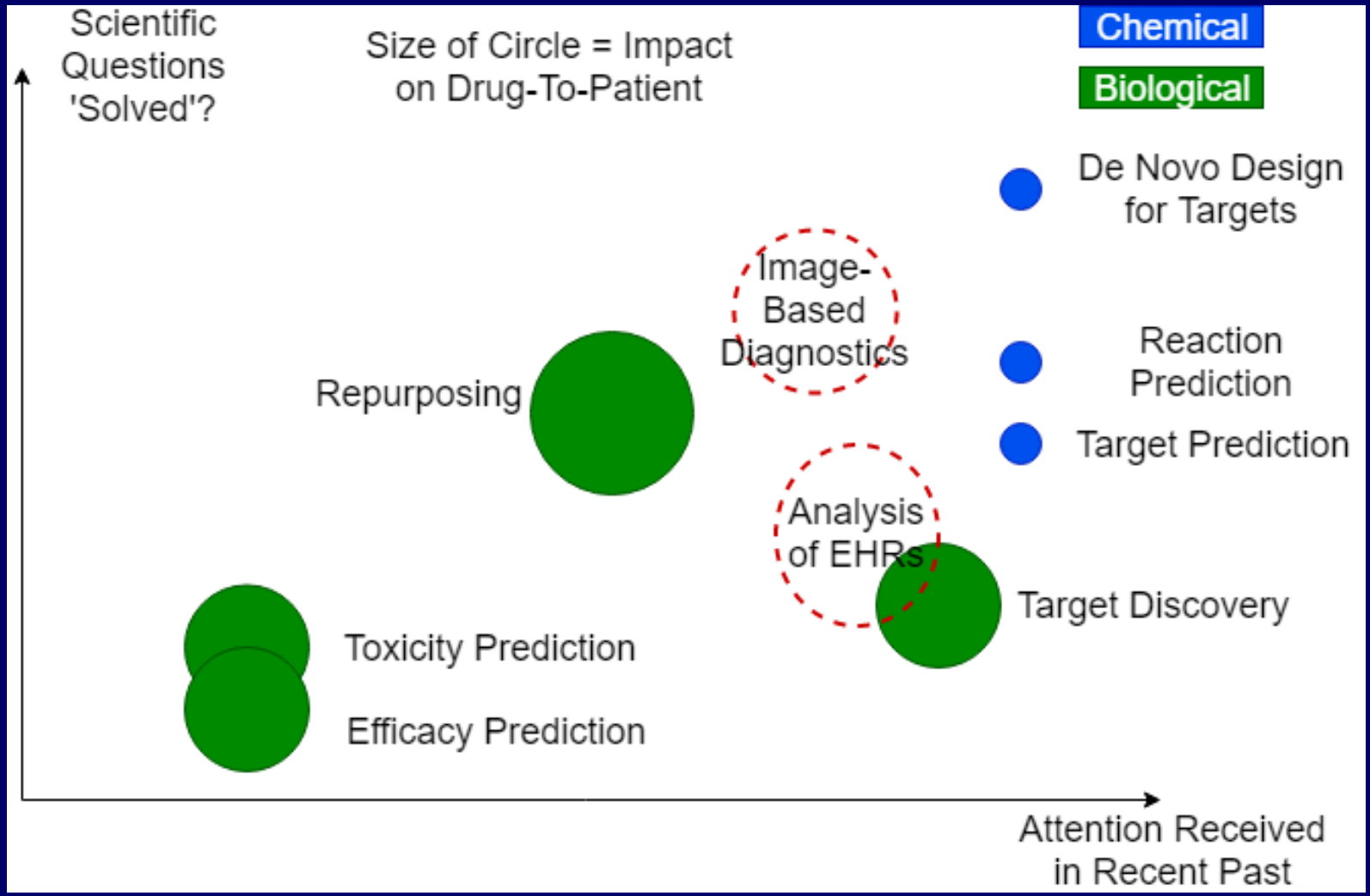
Using bioactivity data for ligand-protein activity modelling '*is relatively possible*'

- On-target bioactivities (links between chemical structure and protein targets) are data-rich, and relatively homogenous
- Hence, generating models for on-target bioactivities is 'possible'
- Can also be used for design (eg multi-target ligands)

BUT:

- Only covers known chemical space
- Suffers from various data biases; normalizing model output is not trivial, etc.
- Labels are still heterogenous
- *In vivo* relevance needs to be established

Hypothesis: ‘AI in drug discovery’ focuses mainly on chemistry (because biology is too tricky)…?



Everyone *will* disagree on the precise location of points

Key point: AI goes where the data is… so we look for the keys where the light is?

Biological data is *painful*

- Data Scientist: **So *does* drug Y cause adverse reaction Z, or not??**

- Response from Pharmacovigilance Department: If we have a patient with this genotype (which is generally unknown) who has this disease endotype (which is often insufficiently defined) who takes dose X of drug Y (but sometimes also forgets to take it) then we see adverse reaction Z … but only in 12% of all cases and only if co-administered with a drug from class C, and then only in males and long-term (Etc.)

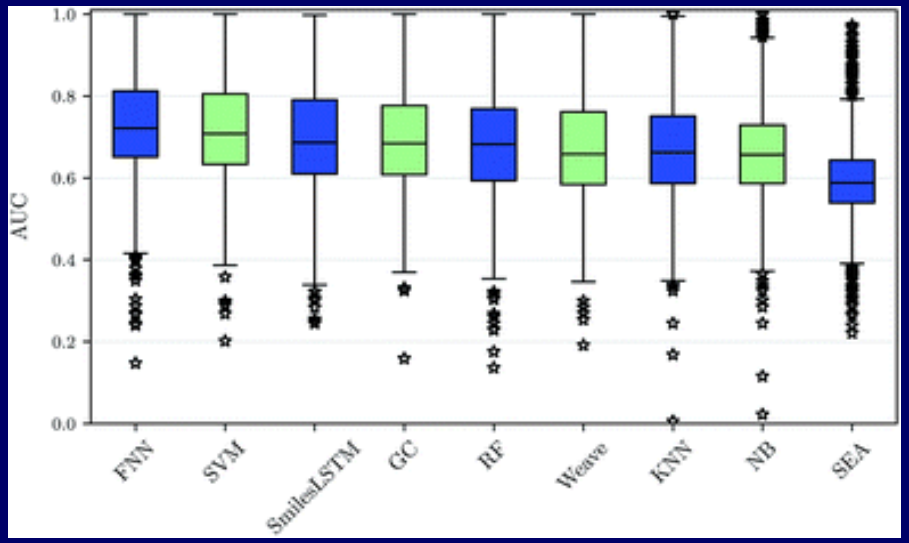
- Analogous for cellular systems (cell line drift, media matter, etc.); animal/histopathology data (is the cage on top or at the bottom? The handler male or female?) etc.

Deep Learning?

- Can work well
- Sometimes works well *numerically*, but it doesn't really address the underlying question
- Is sometimes pushed in a biased ways in publications

There *are* areas in drug discovery where deep learning can work well

Andi Mayr et al. “**Large-scale comparison of machine learning methods for drug target prediction on ChEMBL**”



But trade-off – taking computational time, parameter optimization into account eg for model updates, is it worth it?

- *Statistical significance* is one thing … but does it translate into *practical relevance?*
- "Is your machine learning telling you anything you didn’t already know?“ Anthony Nicholls' slides from 'AI in Chemistry' conference in Cambridge September 2019; put online with Ant's permission: http://drugdiscovery.net/data/cambridge_ai.pdf

Modelling synergy of anti cancer compounds using deep learning

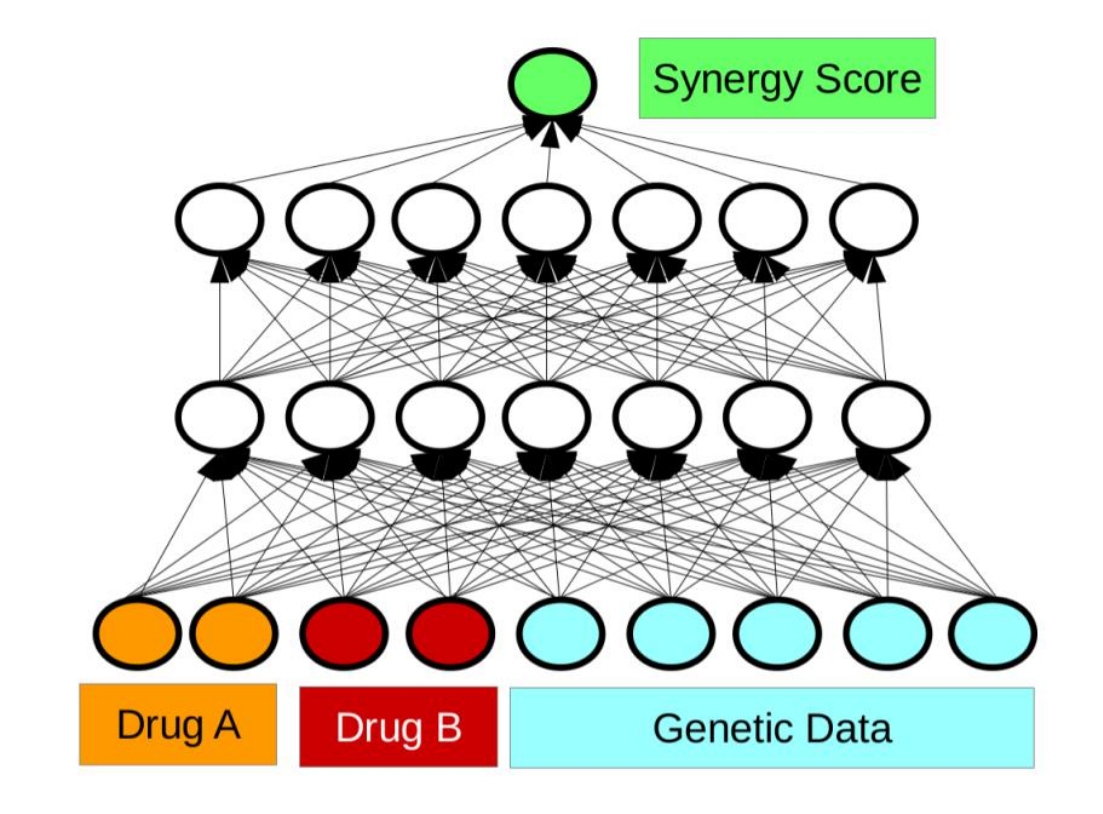
- Sometimes synergy between drugs is desired (in cancer, infectious diseases, …) to ideally improve efficacy/decrease side effects of treatment

- Merck, AZ, NCI ALMANAC, … recently published combination datasets which were can use to model combination effects

- Self-critical evaluation of our work: So does this matter in drug discovery, in practice – in the *real world*?

- Preuer *et al.*, Bioinformatics 2018

Models Used: Deep Neural Networks ('DeepSynergy')



Compared to: median polish, Elastic nets, Random Forest, SVM, Gradient Boosting Regression

DeepSynergy model results: Classification and quantitative model

- Synergy score of 30 as threshold: True Positive Rate 0.55, True Negative Rate 0.95

- *‘1 out of 2 positive synergistic predictions is correct, on average, while 19 out of 20 non-synergistic predictions are also correct, and can be rightly discarded, when looking for synergistic compound combinations’*

- But: Practical relevance? Synergy is dose dependent; and ***does it translate to in vivo situation***….? (Greater question: Do simple endpoints, which we need for AI, really help??)

- **Sometimes we maybe only play a ‘My numbers are higher than yours’ game in the end…**

“You see what you want to see” – biased reporting

Article | Open Access | Published: 08 May 2018

Scalable and accurate deep learning with electronic health records

Alvin Rajkomar ✉, Eyal Oren, […] Jeffrey Dean

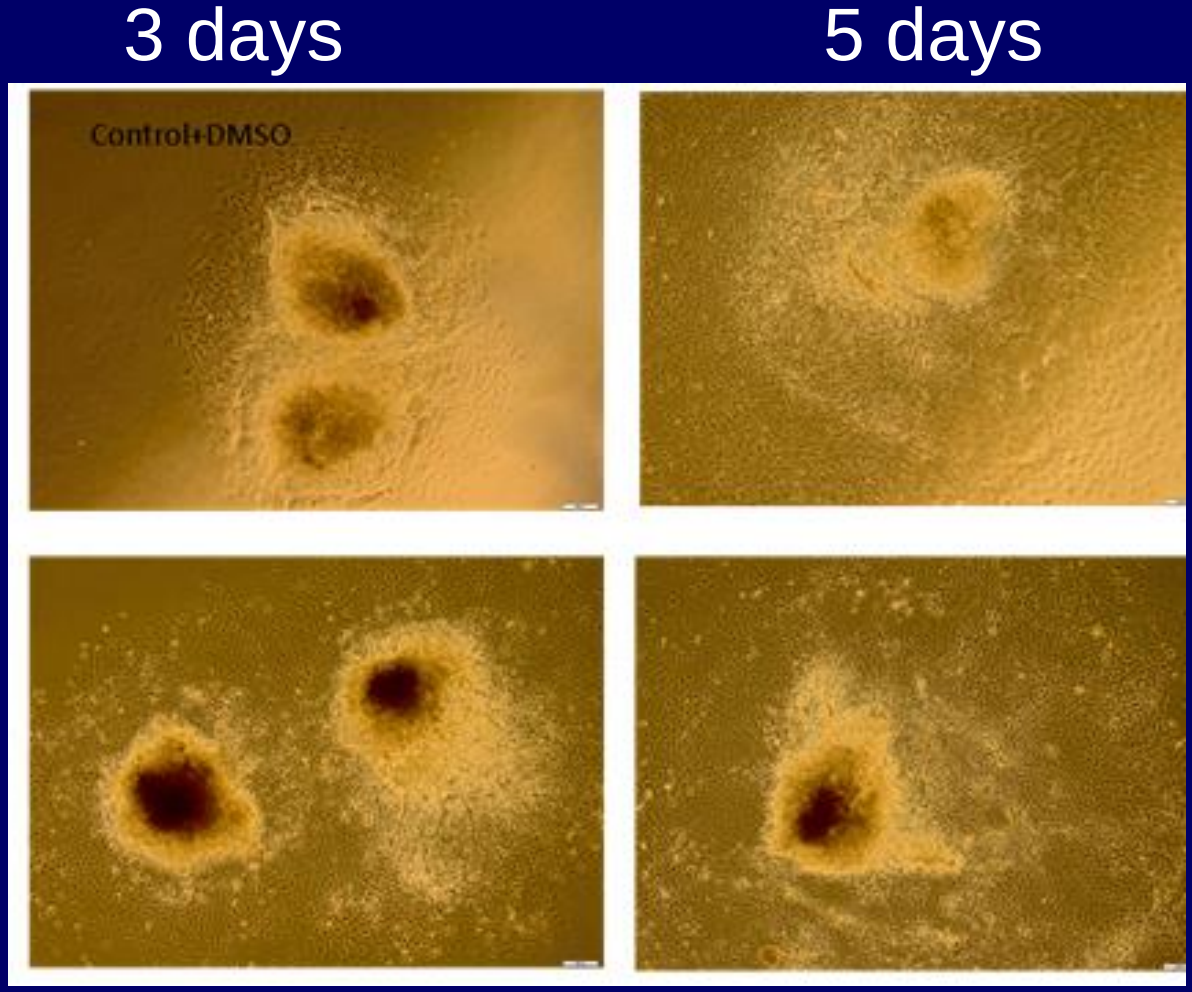
Abstract: “Deep learning models achieved high accuracy for tasks such as predicting: in-hospital mortality (area under the receiver operator curve [AUROC] across sites 0.93–0.94), 30-day unplanned readmission (AUROC 0.75–0.76), prolonged length of stay (AUROC 0.85–0.86), and all of a patient’s final discharge diagnoses (frequency-weighted AUROC 0.90).”

Logistic regression baseline (last page in SI): “For the full feature enhanced baselines, for predicting inpatient mortality at 24 hours after admission, the AUROC was 0.93 (95%CI 0.92-0.95) for Hospital A and 0.91 (95%CI 0.89-0.92) for Hospital B. For predicting unexpected readmissions within 30-days the AUROCs at discharge were 0.75 (95%CI 0.73-0.76) for Hospital A and 0.75 (95%CI 0.74-0.76) for Hospital B. For long length-of-stay at 24 hours after admission, the AUROC was 0.85 (95%CI 0.84-0.85) for Hospital A and 0.83 (95%CI 0.83-0.84) for Hospital B.”

-Omics data is often difficult to ‘model’, but it *can* contain signal!

- Eg repurposing
- Distinguish: ‘One out of many’ selections, or definite predictions for a given molecule (!)
- In our experience eg transcriptomics data often contains sufficient signal for *signal detection* (but, possibly, less so for ‘modelling’)
- ’1/3 of the time nothing happens, 1/3 of the time too much happens, and 1/3 of the time you see something you can use (though you might have already known this beforehand anyway)’

Selected compound induces differentiation of stem cells into cardiac myocytes (by RT-PCR; work with Dr Nasr, Royan Institute, Isfahan)



KalantarMotamedi *et al. Cell Death Discovery* **2016**

Discussion

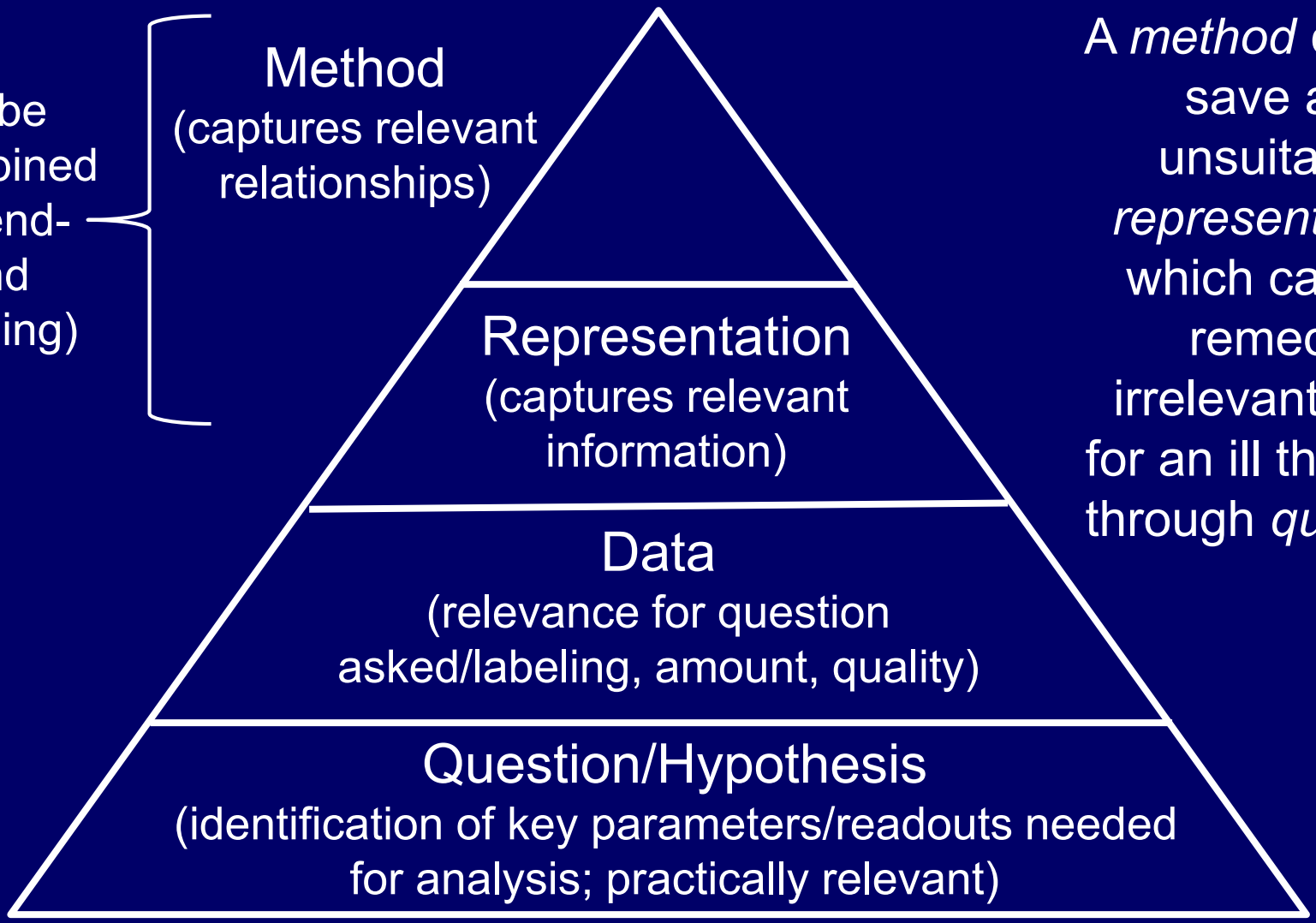
- Our data
- Technical problems with modeI generation
- AI and human beings
- AI and society

Much of the data we generate is generated for the wrong reasons (or in wrong ways)

- Often proxy measures (to reduce cost)
- Irrelevant system/dose/time point
- Often hypothesis-free (‘here we have our pile of data … anyone wants to have a go at it?’) instead of hypothesis-driven
- Often ‘technology push’, instead of ‘science pull’

http://www.drugdiscovery.net/2020/01/21/omics-data-so-where-is-the-signal-please/

First the question, then the data, then the representation, then the method!



Can be combined (eg end-to-end learning)

A *method* cannot save an unsuitable *representation* which cannot remedy irrelevant *data* for an ill thought-through *question*

Is it the method… or is it by chance?

- If a drugs results from the ‘drug discovery pipeline’ it is *the result of a long series of choices*

- Claim: “AI discovers a drug against X!”

- What is responsible?

  Impossible to say!

- Viewpoint A: ‘We don’t have a baseline for control!’
- Viewpoint B: ‘But it worked – look at the compound!’
- Both true at the same time

- Problems: Biased reporting; no baseline control; *focus on trivial wins*

“We cannot validate a model properly”

- Performance and data are related; data usually not sufficiently characterized to put performance into context (numbers only are hence meaningless!)

- ‘Apart from true ***large-scale/diverse prospective validation***, which is often impossible, we cannot have a true idea of model performance’

- Comparative datasets are retrospective… but since (a) they are limited in size (compared to chemical space) and (b) we don’t know underlying distributions (in chemical space) *we will never be able to have a true estimate of model performance!*

- We only play the ‘my number is higher than yours’ game…

The bigger picture: ‘AI’ is where it is due in no small part due to human psychology

- Hype bring you money and fame – realism is boring
- FOMO (‘the others also do it!’) and ‘beliefs’ often drive decisions
- ‘Everyone needs a winner’ (‘*after investing X million we need to show success’)*
- Selective reporting of successes leads to everyone declaring victory (but in reality no one knows what’s actually going on)
- Difficult to really ‘advance a field’ with little real comparison of methods

Summary

- We need to analyse our data (as we did for many years before), absolutely!
- ‘AI’/deep learning is a valuable tool in the toolbox
- The real game changer for translation to patients will come only once we understand biology/biological data better (and generate it, and encode it, and analyse it)
- Currently a lot of computer science-driven approaches, some of which are more applicable in drug discovery than others (real translation is necessary, *but also better experimental design!*)
- Consortia on even larger scale are needed (for targeted data generation, not just sharing what is there already)

Thank you for listening

ab454@cam.ac.uk